Image retrieval by hypertext links

V. Harmandas, M. Sanderson, and M.D. Dunlop
{harmandv | sanderson | mark}@dcs.gla.ac.uk
Computing Science, University of Glasgow
Glasgow G12 8QQ, Scotland.

Abstract

This paper presents a model for retrieval of images from a large World Wide Web based collection. Rather than considering complex visual recognition algorithms, the model presented is based on combining evidence of the text content and hypertext structure of the Web. The paper shows that certain types of query are amply served by this form of representation. It also presents a novel means of gathering relevance judgements.

1 Introduction

Although there have been several approaches to combining *hypertext networks* with *information retrieval* (IR) engines (*[Agosti 96]*, *[Frisse 88]*, *[Croft 93]*, *[Dunlop 93]*, & *[Frei 92]*), there has been little exploitation or testing of these techniques on the *World Wide Web* (or Web). This paper exploits the linked nature of the Web to provide access to images based on an existing model developed for image retrieval from other hypermedia collections. The paper initially develops the model of retrieval and then presents the results of a series of experiments with a large Web based collection.

2 The model

Dunlop *[Dunlop 91]* introduced a model for retrieval of documents by context (as defined by a hypermedia/hypertext) when content based retrieval was not possible (e.g. for non-textual nodes). Hypermedia links are used to calculate an approximation to the content of a non-textual node by using clustering techniques; this approximation, or *representation*, is then treated as the document's content for use by the retrieval system. This section gives an overview of this model, describes how it is extended to work with Web based collections and describes the main limitations of the model.

2.1 Basic model

In a hypermedia collection, links can be used to calculate representations for non-textual nodes that permit direct retrieval of these nodes by textual query. The textual nodes linked to a non-textual node can be considered as forming a cluster, see Figure 1. Cluster description techniques can then be applied, in order to calculate a representation and establish the overall content of the documents

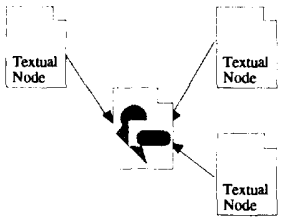

Figure 1. A non-textual node linked to by textual nodes.

forming the cluster. This representation can be subsequently assigned to the non-textual node, giving it a retrieval content equal to the combined content of the nodes connected to it.

The representation of a non-textual node can be calculated by considering each linked document L as a standard N-dimensional vector, where N is the number of index terms in the document base. The centroid in the N-dimensional space of the points representing the documents in the cluster can then be calculated using the following formula *[Dunlop 93]*: The formula defines a level one cutoff, where only the immediate neighbours of a non-textual node in the hypermedia network are considered. Such nodes are said to be connected to the non-textual node by a *one step link*.

$$W_{di} = \frac{\sum_{L \in C_d} L_i}{|C_d|}$$

where

i ranges between 1 and N;

W_{di}= cluster based weight of term i of document d;

L_i= the weight of term i in the document L;

C_d= cluster of documents linked to, and from, document d.

To test the validity of the approach, two experiments were carried out using the CACM collection *[Dunlop 93]*: containing text documents with citation information that, in effect, forms a hypertext structure between some of the documents. The experiments initially calculated the representation of each document by traditional statistical analysis of the words in the document and then a second representation for each document based on the cluster technique using the text of documents citing, and cited by, it. The first experiment showed that the cluster technique assigned representations were considerably closer to the original representations than representations based on randomly generated citations (23% similarity as opposed to 4%). The second experiment showed that citation-

SIGIR 97 Philadelphia PA, USA

based cluster representations provided approximately 70% of the retrieval effectiveness of directly indexing each record's content. These results confirmed that these representations are of suitable quality for use in retrieval (particularly when no content information is available). Later, Savoy *[Savoy 96]* presented results that showed that the CACM collection's citation structure could be used to improve the retrieval of textual nodes by partly including text from linked documents.

Although the results obtained from these experiments are encouraging, nonetheless they were based on a pseudo-hypertext network. It can be argued that the citations and the links in a hypermedia document base share some properties. For example, they are both the result of some judgement on the relatedness of the documents, and there is no single definition of the relationships between the two connected nodes. However, citation links constitute only one type of link that can be encountered in real hypermedia, other types being, for example, structural, relational, etc. Section 2.2 discusses a Web specific version of the algorithms that the experiments presented in this paper are based on.

The algorithm described above can be extended so as to take into account not only the immediate neighbours of a node in the hypermedia network, but also all nodes that can be reached from the node by following at most two links (a *two step link*), and thus consider more of the context of the node (e.g. to include all the textual documents shown in Figure 2). This results in the following level

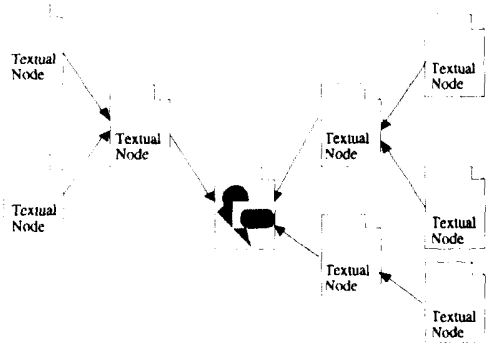

Figure 2. A non-textual node and the nodes accessible from it by one and two step links.

two cut-off formula:

$$W_{di} = \frac{\sum_{L \in C_d} L_i}{|C_d|} + k\frac{\sum_{L \in C_d} L_i}{|C'_d|}$$

where

i ranges between 1 and N;

W_{di}= cluster based weight of term i of document d;

C_d= cluster of documents linked to, and from, document d;

k is a constant, $0 \le k \le 1$, defining the relative strength of the more remote neighbours;

$$C'_d = \bigcup_{i \in C_d} C_i - C_d - d .$$

The model described above presupposes a hypermedia network, which will provide the textual and non-textual interconnected nodes. Dunlop already implemented this model using a hypermedia version of the British Highway Code. His experiments concentrated on evaluation of the model's utility when combined with browsing, rather than on its retrieval effectiveness. This paper investigates the use of the model on the Web. There are two major reasons supporting this decision:

- The Web is a real hypermedia network, and currently the only hypermedia structure used by millions of users;
- its diversity in information, variety of media, and complexity of interconnections make it an ideal test bed to verify the validity of the model.

2.2 Web specialisations

Given a hypermedia collection consisting of textual and non-textual nodes interconnected via links, the steps required to provide the non-textual nodes with a representation are as follows:

- Firstly, all the non-textual nodes in the collection are identified;
- for each of these non-textual nodes, all the textual nodes that are linked to them, by means of one and two step links (nodes that are reached from the current node via one or two hypertext links respectively), are also identified;
- the content of these nodes is used to form a new node, which is the representation of the non-textual node;
- finally, the collection of representations is indexed, using standard IR techniques (including stemming, stop word removal, and term weighting).

With this index, the documents corresponding to the non-textual nodes can be retrieved in the same fashion as any standard text document. It is the task of the retrieval system's interface to associate the documents retrieved with their corresponding non-textual nodes and present them to the user. Although the model is designed to be general enough to permit its application on any type of media (e.g. images, sound, and video), this paper concentrates on images for two reasons:

- the Web has an abundance of images, and thus, creating an image collection would be relatively straightforward and the results would be of widespread applicability;
- secondly, focusing on a single media avoids possible performance differences between media interfering with results.

The first stage of indexing a Web collection is to identify the documents in the collection and their links. A Web crawler/robot *[Cheong 95]* was used to scan a set of image collections and store, for each image, the text of pages linked to that image via one and two step links.

Next an approach to extracting the text from nodes to create suitable representations must be developed. During this development of the model presented earlier for Web use, two specific issues concerning the Web structure became evident: some pieces of text within a page may be more important for indexing than others, and some images retrieved by a simple application of the model are unlikely to be valid search results. The following sub-sections deal

with these Web specialisations of the Dunlop model (presented in Section 2.1).

2.2.1 Extracting suitable text

A Web page is the result of viewing an *HTML* document using a Web browser and typically consists of formatted text, images, fill-out forms, tables, anchors to other parts of the same document, and links to other HTML documents[1]. Images are referenced in a page via a URL link and are displayed in one of two ways: either as in-line images displayed in the page, or via a link to an image file.

At the initial phase of the model implementation, a decision was made to split the text of a page into independent sections according to its position in respect to the URL link to an image. Text adjacent to the link was thought to form a more accurate representation than the text of the whole page, as it is usually the practice that an image is briefly explained by means of a caption. Figure 3 shows an

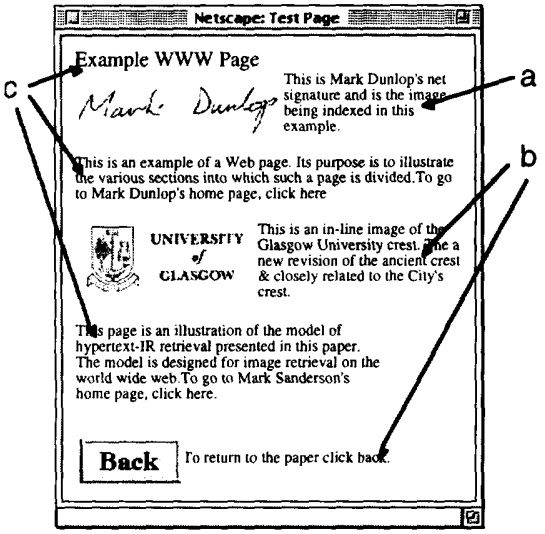

Figure 3. Example HTML page with sections indicated.

example HTML page where the image of the signature is to be indexed by the surrounding text. The text is broken up into three sections:

- *image caption* (a): the caption of the image being indexed. Defined as the text after the image's URL until the end of the paragraph (i.e. indicated by '<p>' or '<br>' tags) or until a link to another image is encountered;
- neighbouring image captions (b): the captions of other images within the page;
- one step link text (c): the text of the page, excluding that contained in sections (a) and (b).

In addition, there is one extra section defined:

- two step link text (d): the full text of all pages connected via two-step links to the image or image file: in this example, the text of all documents linked to the page in Figure 3.

1. A detailed description of the HTML language and be found in December and Ginsburg *[December 95]*.

Splitting document texts according to these definitions yields, for each of the images in the collection, a text document consisting of the four sections described, thus producing a pure text collection representing the images referenced in the Web collection.

This hypothesis on text segmentation and the quality of automatic segmentation schemes required testing together with the main hypothesis of the representation method. To achieve this, the collection was indexed and matched to queries using SMART v11.0 *[Buckley 85]* with the results presented in Section 3.3.

2.2.2 Removing unsuitable images

It is usually the case that the images referenced in a page are semantically related to the textual content of that page, with the exception of functional images (see Figure 4 shows examples of

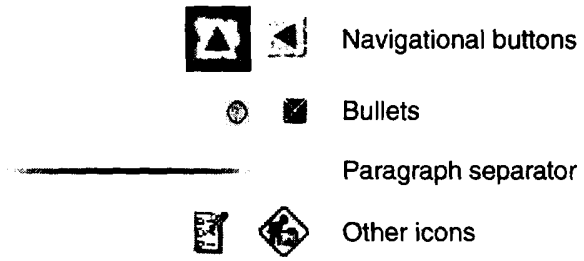

Figure 4. Example functional images.

such images):

- images denoting that the page is new or under construction;
- images that are used as paragraph separators, indicators of list elements;
- navigational buttons, used to help the user navigate to the next, previous, or home page, or even to the top or bottom of the same page;
- icons used to enhance the presentation of the page and/or make it more readable.

Although functional images are used to give structural and navigational information to the reader, there is no direct way of preventing a system from retrieving these images as relevant to a query as HTML does not distinguish them from other images. However, it was decided that a way should be found to eliminate, or at least, reduce the possibility that these images are retrieved, since they are unlikely to be the legitimate results of a user's search. One clue is that these images tend to be linked with more pages than non-functional images, since these images are usually navigational or readability enhancement images that can be found in many pages. Furthermore, since the set of pages forming a collection are part of a specific network, it is highly probable that the same images are shared by these pages, primarily for reasons of presentation uniformity. Figures 5 & 6 show the number of images (expressed as a percentage) that have a particular number of links to them. Measurements were made over a number of Web sites for around 2,500 images. From these figures it can be seen that functional images are always linked to by more than two pages, with almost half of them being linked to by more than five. Non-functional images, are linked to by only one page over 90% of the time.

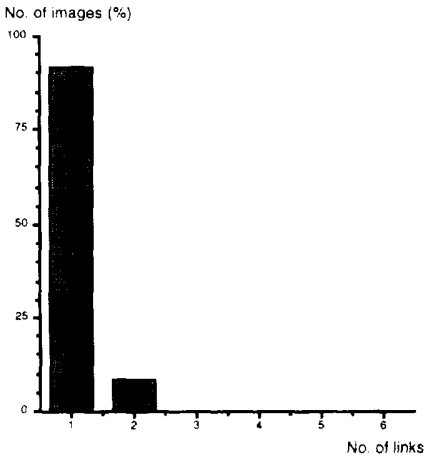

Figure 5. Percentage of non-functional images that have a particular number of links to them.

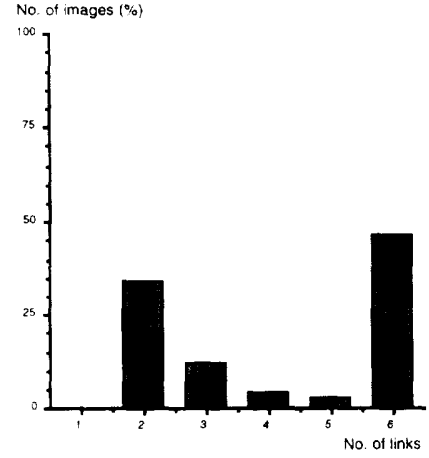

Figure 6. Percentage of functional images that have a particular number of links to them.

A reasonable way, therefore, to reduce the possibility of functional images being retrieved, was to weight them inversely to the number of pages linking to them and let the system retrieve them, but with a low relevance score. The application of this weighting function on the images retrieved was found to push most of the functional images down the ranking list without significantly affecting the position of other images.

2.3 Limitations of the model

Dunlop presents some limitations of his model that are mainly connected with the quality and quantity of the links available in a given document base. He restricts his model to document bases that have "a reasonable ratio of indexable (i.e. textual) to non-indexable nodes" and recommends at least two links per image. Moreover, the model benefits from an increased number of links (as long as the number of nodes used to calculate a cluster representative remains a small subset of the collection) but appears to degrade smoothly (so long as some links are available). The implications of the low link levels shown in Figure 5 are addressed in Section 3. As will be seen, these limitations affected the choice of Web sites to use for experimentation.

Another limitation expressed by Dunlop is imposed by the quality of the links. If the links are to very similar nodes, then the representations will more precisely describe the content of the cluster. Conversely, if the links are to loosely related nodes, then the cluster representative calculations will yield a less accurate description of the non-textual node, and thus the retrieval effectiveness will decrease. Clearly this translates directly to Web-based collections and will be a critical factor in the experimental results presented in this paper.

2.4 Existing web based image searchers

Recently, some of the commercial Web search engines, such as Lycos *[Lycos]* and Hotbot *[Inktomi]*, started supporting image retrieval using a form of text based representation. Presumably for commercial reasons, it is not possible to find how these representations were generated or how effective the retrieval methods were. A short investigation was undertaken to discover the general form of the representations used. This was done through the submission of queries to these systems and the examination of the pages retrieved.

From the investigation, it would appear that Lycos represents images by a small amount of text, often only a few words. This representation appears to severely limit the range of images retrieved for a particular query. Hotbot's representation of an image on a Web page is all the text on that page. It is believed that there is no further sophistication to this representation, but within the confines of the investigation, it was not possible to determine this for certain[2].

Additionally, there is the WebSeer retrieval system by Frankel et. al. *[Frankel 96]* which uses a combination of surrounding text and image analysis techniques to form a representation of an image. Surrounding HTML text is broken up into sections, though the section definitions used are different from those described in Section 2.2.1. It would appear that no experiments have been conducted to measure the retrieval effectiveness of the system. Instead, most effort has been concentrated on the building and testing of the image analysis techniques:

- determining if an image is a photograph or not;
- if an image is in colour or in black and white;
- locating faces within an image.

3 Experimental design

In order to evaluate the effectiveness of an IR system based on the Dunlop model, a test collection was created. Sparck Jones and Van Rijsbergen *[Sparck Jones 76]* suggest that the ideal collection should:

- be large, i.e. contain no less than 2,000 documents (in some cases, even 10,000 documents could be needed), while queries should be more than 250;
- exhibit on the one hand variety with respect to content, type, source, etc., and on the other homogeneity.

2. This investigation was carried out in early 1997, it is of course possible that the methods used by these search engines have subsequently changed.

This section describes the process of building the collection and then the results of various experiments with the collection.

3.1 Document gathering

The Web is a good source of documents that hold to the requirements of Sparck Jones and Van Rijsbergen. It consists of a large number of autonomous 'Web sites' that are disparate when compared. Most, however, can be categorised into genres: homogeneous groups of sites that cover the same general topic. It was decided to select one particular genre and use sites belonging to that genre as a document collection. It was decided to use sites belonging to the genre of on-line art galleries. There are four main reasons that justified this decision:

- first, the Web hosts a broad variety of art sites, in the form of either on-line exhibitions, or virtual galleries;
- secondly, these sites are usually populated by many images (their number ranges from 100 to 1,500), therefore, the goal of creating a collection of no less than 2,000 documents could be achieved without the collection being fragmented into too many independent sub collections.
- most of these sites have large portions of text associated with the images, either in the same page, or in neighbouring pages, therefore, each image could be adequately represented in terms of the model developed in Section 2;
- finally, it was anticipated that it would be fairly easy to find subjects with suitable knowledge and interest to query an art-related collection.

The image collection was created out the pages of seven unrelated Web art gallery sites, containing art from the renaissance to the 20th century. In total, the collection was composed of 2,583 images and 6.5Mb of HTML text. The selection of the Web sites was based on three criteria derived jointly from the criteria of Sparck Jones and Van Rijsbergen and the limitations expressed by Dunlop:

- each of them should contain more than 100 images;
- the overlapping of images between the collections should be minimal;
- they should have an extensive hypertext structure, therefore, collections containing a large number of images in only one or two pages, were excluded.

Obviously these criteria introduce a bias into the collection with respect to 'normal' Web sites, it was felt that this was justified as the main objective of the experiments was to show whether or not the representation of images using this model would work at all.

3.2 The queries and relevance assessments

The second task towards the creation of a test collection was the assembly of the queries and the identification of all images relevant to each query. Queries were provided by four people: two post-graduate students of Art History, and two students in other disciplines, but with a interest in art and a broad knowledge of art-related subjects. The four were informed of the type of the collection and the kind of images that it contained (i.e. the kind of artistic movements that the sub collections covered), and were then asked to form queries that they would actually submit to a retrieval system, should they have a real information need relating to the collection's content. The queries obtained in this way were then processed, so that, first, they did not have the same subject, and second, they were not clearly out of context. This process yielded a set of 14 queries, shown in Figure 7.

1	Pop art of the 1960s
2	Dutch baroque
3	European baroque
4	Lucian Freud
5	American paintings of the 1930s
6	Landscapes by Monet, Manet, and Sisley
7	Caravaggio
8	Wassily Kandinsky
9	Goya around 1795
10	German gothic
11	Raphael's major syntheses from 1506 to 1508
12	Mondrian Piet and Purism
13	Matisse
14	Byzantine icons

Figure 7. Queries used in retrieval experiments.

Undeniably the size of the query sample is small and far from satisfies the requirements of an ideal test collection. However as already stated, the purpose of this work was to establish the validity of the Dunlop model for this type of collection, with a view to conducting more detailed experiments later. Therefore, it was decided this number of queries was satisfactory for these initial purposes.

3.2.1 Relevance assessments

The relevance assessments for each query were made by the person that generated that query. Rather than manually assess all images in the collection for relevance, the structure of the Web sites was used to guide the assessor in choosing a sub-set of images to assess. For example, if a query was about the Early Renaissance, then the assessor would look for relevant images in pages that were classified under only this or similar titles. It was decided that this targeting approach was valid as the creation of the hypertext structure within the Web sites had involved manual decisions about where in this structure images should be placed. In effect these decisions were a form of relevance assessment on the images in relation to the categories embodied within the structure. By focusing on potentially relevant images, the time taken to make the relevance judgements was significantly reduced. Clearly, this exploitation of the Web sites' structure would only be of use for queries that reflected the classifications contained in the structures. Fortunately, this was the case for the queries generated for these experiments.

The relevance assessors identified a total of 307 relevant images for the 14 queries, on average, 22 images per query.

3.3 The experiments

As was outlined in Section 2.2.1 the representation of each image is by four text sections:

- the image caption: a;

- captions of other images in the text: b;
- body text of documents with a one step link to the image (without sections a and b): c;
- text of documents with two step links to the image: d.

3.3.1 Measuring the utility of the sections

The initial experiment involved measuring the retrieval effectiveness of the IR system in four different configurations: each time using just one of the identified text sections in isolation. A graph showing the results of this experiment is shown in Figure 8. As can

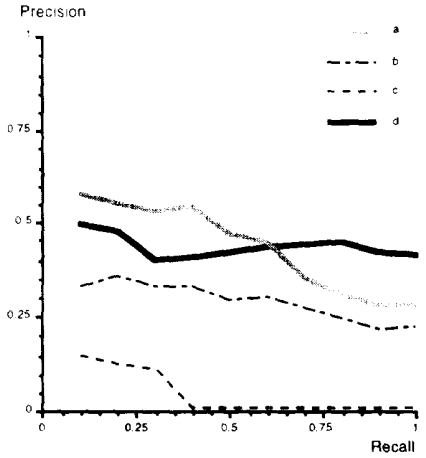

Figure 8. Results of four experiments to discover the utility of representing images by each of the four types of text section.

be seen, effectiveness is highest for images represented by section a and by section d: the image caption text and the 'two step link' text respectively. Perhaps surprisingly, representation by the text of other image captions on the HTML page (section b) also produces relatively high effectiveness in contrast to use of the body text of the HTML document (section c) which has the lowest utility in terms of effectiveness. Reasons for this result are not clear, one possibility is that the HTML pages indexed in this collection consisted of almost nothing but image captions and therefore, there was little text identified as section c.

3.3.2 Finding the best combination of sections

For these experiments, the sections were assigned weights that reflected their significance in the representation of an image. Measurement of the retrieval effectiveness of the IR system using all possible combinations of section weights, ranging between one and four, was performed. The 256 sets of *recall/precision* (RP) figures resulting from these experiments were each reduced to a single figure using f_{max}: a statistic based on the e measure taken from Van Rijsbergen *[Van Rijsbergen 79]* and defined by Sanderson *[Sanderson 96]*. Not unsurprisingly, the best combination of section weights reflected the results of Section 3.3.1: emphasising sections a and d resulted in the highest effectiveness and emphasising sections b and c resulted in the lowest. Figure 9 shows retrieval effectiveness resulting from the best and worst section weight assignments found: the best was in assigning the weights 4, 1, 1, 3 to sections a, b, c, d respectively; the worst was in assigning 1, 2, 4, 1.

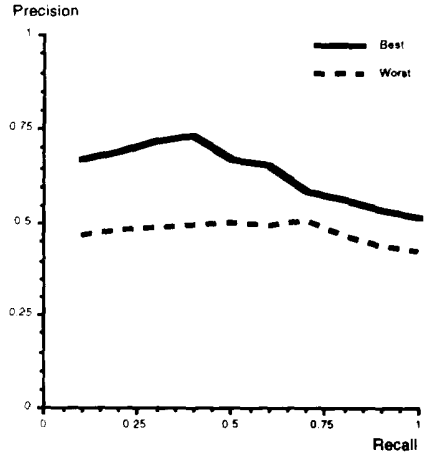

Figure 9. Comparison of the worst and best assignments of weights to the four text sections.

3.3.3 The shape of the R/P graphs

Through out these experiments, it was noted that the RP figures were somewhat unusual when compared to other IR experiments. It is generally the case that a RP graph displays a monotonically decreasing line. In these experiments, the lines decrease less than might be expected and on occasion increase at higher recall levels. It is believed there are two reasons for this.

- The R/P graph resulting from an individual query is most often an uneven line quite unlike that derived from the results of a large number of queries. Because of the small number of queries generated for this collection, the unevenness of individual queries may not have been averaged out.
- Upon analysis of the R/P figures of individual queries, it was found that for many of the queries all relevant images were retrieved near the top of the document ranking. In other words, for the IR system, the collection contained a high number of easy queries. It is this aspect of the queries that explains the 'flatness' of the R/P graphs.

The only way to change this situation, will be to generate more queries for the collection. Despite these concerns, however, it is believed that the results presented here do provide a good indication of the relative utility (in terms of retrieval effectiveness) of the different text sections used to represent the images.

3.4 Experiments with other types of query

The queries used in the retrieval effectiveness experiments were all concerned with artists or art movements, and from the experiments it would appear that the representation of the images worked well for these queries. Some additional ad hoc tests were performed to examine what other types of query could be best served by this representation. Those that worked well were as follows:

- queries for images containing certain objects, for example images containing a mountain; or apple, fruit, and a table (see Figure 10);
- queries for types of images such as a street scene or a still life;

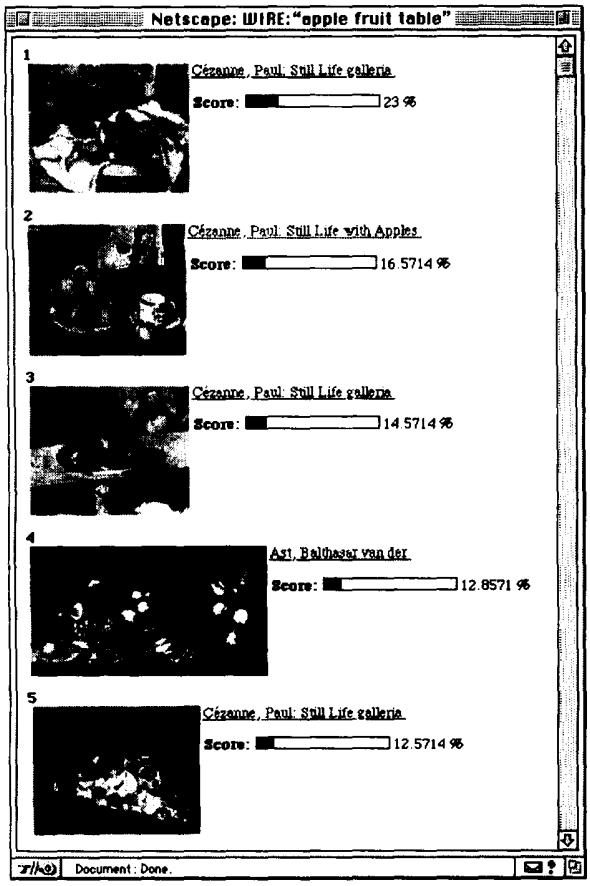

Figure 10. Screen shot of a sample retrieval.

- queries for a particular painting, for example "The Scream by Munch".

Those that did not work were:

- abstract queries for images conveying a certain emotion or intent;
- queries on the superficial aspects of images, for example images of a certain size or those that contain an amount of a certain colour.

4 Conclusions and future work

In this paper an existing retrieval model (from *[Dunlop 93]*) for representing images by their surrounding text was successfully applied to a collection of 'art gallery' Web sites. The retrieval effectiveness of this model was measured by constructing a test collection of approximately 2,500 images. One contribution of this work was the use of a technique to aid the process of gathering relevance judgements: this traditionally time consuming process was significantly speeded up by exploiting the existing hypertext structure of Web sites to target areas of those sites that might contain images relevant to a query.

The main contribution of this work was to show through experimentation that the Dunlop model could be used as an effective means of representing images for a number of query types. This, despite that fact that the text used in the representation was not written explicitly for the purposes of retrieval. The representation of an image was composed of sections each of which was assigned a weight that indicated that section's significance in a retrieval. Through experimentation, it was found that by weighting higher the image caption section and the two step link section, retrieval effectiveness could be improved.

In future work for this project, it is planned that the collection will be expanded and a more diverse set of queries be generated. Dunlop also proposed a number of extensions to his model and these could be examined. One possible avenue of investigation would be to consider the direction of the links when calculating the representation. For example, in the case of bi-directional links, links in one direction could be ignored, links from and to the node could be considered as equal, or the one direction of links could be weighted over the other.

Ultimately, the aim of this work is to integrate the text based representations used here with content based representations. For example the recent work of Fleck, Forsyth and Bregler *[Fleck 96]*, whose retrieval system analyses the content of images to locate images of people in a particular state. By integrating the two representation schemes into a single retrieval system, it would be hoped that a wider range of query types could be supported and retrievals from those queries be of a greater accuracy.

5 Acknowledgements

The work of this thesis was carried out by Vassilis Harmandas as partial fulfilment of his M.Sc. in Advanced Information Systems (AIS) at the Department of Computing Science at the University of Glasgow. It was funded by the Alexander S. Onassis Public Benefit Foundation. Supervision of the project was by Mark Sanderson and Mark Dunlop. Thanks are due to Evi Athanasekou, Kalliopi Koundouri, Irene Marinos, and Nina Papadoulaki whose relevance judgements and queries made the experiments possible.

6 References

Agosti 96
Agosti M., and Smeaton A. (1996). *Information Retrieval and Hypertext.* Kluwer Academic Publishers, The Netherlands.

Buckley 85
Buckley C. (1985). *Implementation of the SMART Information Retrieval system.* Department of Computer Science, Cornell University, TR 85-686.

Cheong 95
Cheong F.C. (1995). *Internet agents: Spiders, wanderers, brokers, and bots.* Macmillan Publishing, USA.

Croft 93
Croft W.B., and Turtle H.R. (1993). *Retrieval strategies for hypertext.* Information Processing & Management, 29(3), pp. 313-324.

December 95
December J., and Ginsburg M. (1995). *HTML & CGI Unleashed.* Sams Publishing, Indianapolis.

Dunlop 91
Dunlop M.D. (1991). *Multimedia Information Retrieval*, Ph.D. Thesis. Computing Science Department, University of Glasgow, Report 1991/R21.

Dunlop 93
Dunlop M.D., and Van Rijsbergen C.J. (1993). *Hypermedia and free text retrieval.* Information Processing & Management, 29(3), pp. 287-298.

Inktomi
Inktomi Corp., HotBot™ Search, http://www.hotbot.com/.

Fleck 96
Fleck M., Forsyth D., and Bregler C. (1996). *Finding Naked People*, Proceedings of 1996 European Conference on Computer Vision, Volume II, pp. 592-602.

Frankel 96
Frankel C., Swain M. J., and Athitsos V. (1996). *WebSeer: An Image Search Engine for the World Wide Web*, University of Chicago Technical Report TR-96-14.

Frei 92
Frei H.P., and Stieger D. (1992). *Making use of hypertext links when retrieving information.* Proceedings ACM-ECHT'92, Milan, Italy, pp. 102-111.

Frisse 88
Frisse M.E. (1988). *Searching for information in a hypertext medical handbook.* Communications of the ACM, 31(7), pp. 880-886.

Lycos
Lycos Inc., Lycos™ Search, http://www.lycos.com/.

Van Rijsbergen 79
Van Rijsbergen C.J. (1979). *Information retrieval* (second edition), in London: Butterworths.

Sanderson 96
Sanderson M. (1996). *Word Sense Disambiguation and Information Retrieval.* PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK.

Savoy 96
Savoy J. (1996). *An extended vector-processing scheme for searching information in hypertext systems.* Information Processing & Management, 32(2), pp. 155-170.

Sparck Jones 76
Sparck Jones K., and Van Rijsbergen C.J. (1976). *Progress in documentation.* Journal of Documentation, vol. 32(1), pp. 59-75.